# Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence (Berger & Sellke 2012)

Ryan Giordano (for Broderick ReadStat)

Sep 29th, 2021

Massachusetts Institute of Technology

# Outline

We'll be comparing (frequentist) p-values with Bayesian model selection in a particularly simple setting: testing a point null against a composite alternative with a scalar parameter.

- What is a p-value?
- What is Bayesian model selection?
- Are they commensurable? (No)
- What now?

# What is a p-value?

What is a p-value?

# What is a p-value?

What is a p-value?

## It is garbage.

This is the offical position of the American Statistical Association.

# What is a p-value?

What is a p-value?

## It is garbage.

This is the offical position of the American Statistical Association.

Its executive director writes:

> *"Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true. ... Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof). Don't. Don't. Just...don't." [Wasserstein et al., 2019]*

# What is a p-value?

What is a p-value?

## It is garbage.

This is the offical position of the American Statistical Association.

Its executive director writes:

> *"Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true. ... Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof). Don't. Don't. Just...don't." [Wasserstein et al., 2019]*

Today we will beat this horse a little more.

Whiteboard

# Table: Fixed Normal prior

*Table 1. $Pr(H_0 \mid x)$ for Jeffreys-Type Prior*

| | | *n* | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *p* | *t* | *1* | *5* | *10* | *20* | *50* | *100* | *1,000* |
| .10 | 1.645 | .42 | .44 | .47 | .56 | .65 | .72 | .89 |
| .05 | 1.960 | .35 | .33 | .37 | .42 | .52 | .60 | .82 |
| .01 | 2.576 | .21 | .13 | .14 | .16 | .22 | .27 | .53 |
| .001 | 3.291 | .086 | .026 | .024 | .026 | .034 | .045 | .124 |

# Table: All priors

*Table 4. Comparison of P Values and $\underline{Pr}(H_0 \mid x, G_A)$ When $\pi_0 = \frac{1}{2}$*

| *P Value (p)* | *t* | $\underline{Pr}(H_0 \mid x, G_A)$ | $\underline{Pr}(H_0 \mid x, G_A)/(pt)$ |
|---|---|---|---|
| .10 | 1.645 | .205 | 1.25 |
| .05 | 1.960 | .128 | 1.30 |
| .01 | 2.576 | .035 | 1.36 |
| .001 | 3.291 | .0044 | 1.35 |

# Table: Symmetric priors centered at $\theta_0$

*Table 5. Comparison of P Values and $\underline{Pr}(H_0 \mid x, G_S)$ When $\pi_0 = \frac{1}{2}$*

| *P Value (p)* | *t* | $\underline{Pr}(H_0 \mid x, G_S)$ | $\underline{Pr}(H_0 \mid x, G_S)/(pt)$ |
|---|---|---|---|
| .10 | 1.645 | .340 | 2.07 |
| .05 | 1.960 | .227 | 2.31 |
| .01 | 2.576 | .068 | 2.62 |
| .001 | 3.291 | .0088 | 2.68 |

# Table: Unimodal symmetric priors centered at $\theta_0$

*Table 6. Comparison of P Values and $\underline{Pr}(H_0 \mid x, G_{US})$ When $\pi_0 = \frac{1}{2}$*

| *P Value (p)* | *t* | $\underline{Pr}(H_0 \mid x, G_{US})$ | $\underline{Pr}(H_0 \mid x, G_{US})/(pt^2)$ |
|---|---|---|---|
| .10 | 1.645 | .390 | 1.44 |
| .05 | 1.960 | .290 | 1.51 |
| .01 | 2.576 | .109 | 1.64 |
| .001 | 3.291 | .018 | 1.66 |

# Table: Normal priors centered at $\theta_0$

*Table 7. Comparison of P Values and $\underline{Pr}(H_0 \mid x, G_{NOR})$ When $\pi_0 = \frac{1}{2}$*

| *P Value (p)* | *t* | $\underline{Pr}(H_0 \mid x, G_{NOR})$ | $\underline{Pr}(H_0 \mid x, G_{NOR})/(pt^2)$ |
|---|---|---|---|
| .10 | 1.645 | .412 | 1.52 |
| .05 | 1.960 | .321 | 1.67 |
| .01 | 2.576 | .133 | 2.01 |
| .001 | 3.291 | .0235 | 2.18 |

# Figure

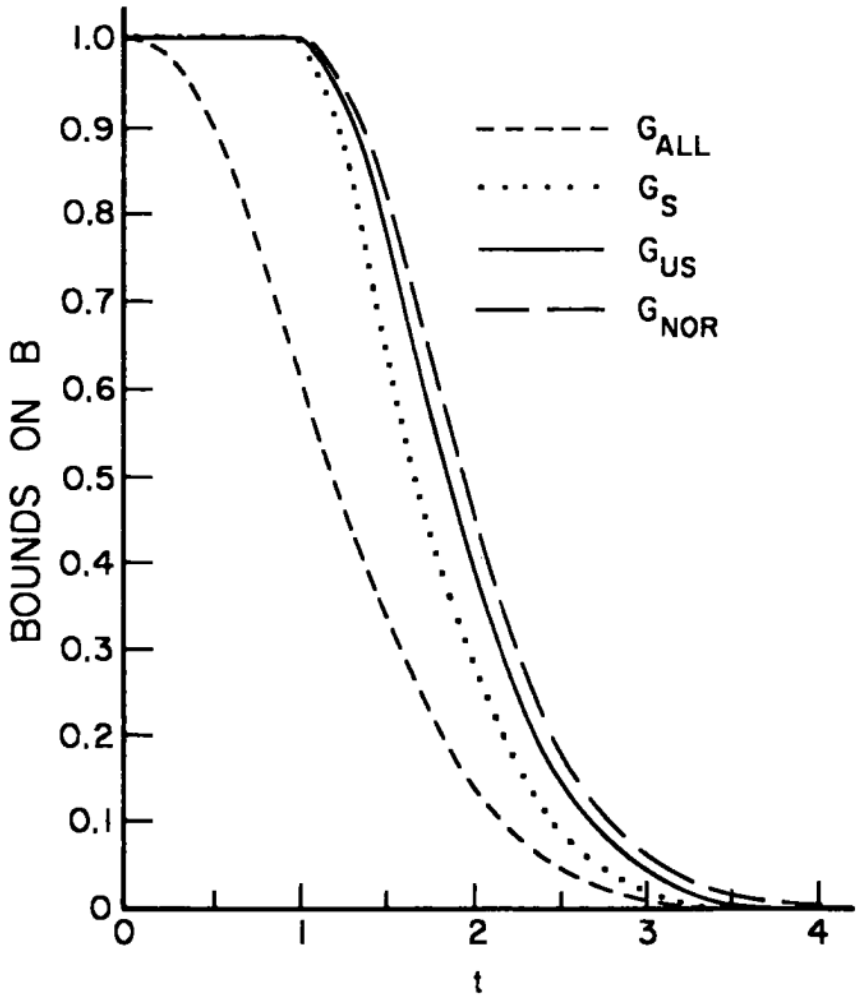


*Figure 3. Values of $\underline{B}(x, G)$ in the Normal Example for Different Choices of G.*

# Figure

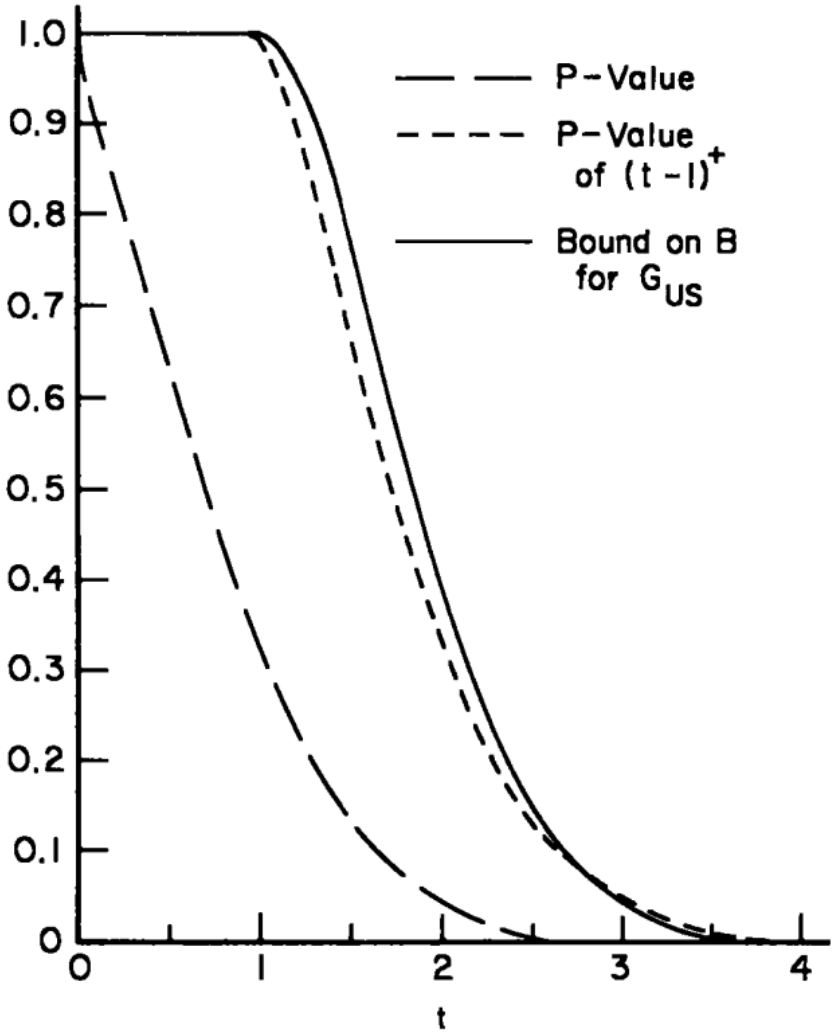


*Figure 4. Comparison of $\underline{B}(x, G_{US})$ and P Values.*

# Table for Cauchy

#notalldistributions

*Table 8. $\underline{B}$ and $\underline{Pr}$ for a Cauchy Distribution When $\pi_0 = \frac{1}{2}$*

| *P Value (p)* | $\lvert x \rvert$ | $\underline{B}(x, G_{US})$ | $\underline{Pr}(H_0 \mid x, G_{US})$ | $\underline{B}(x, G_C)$ | $\underline{Pr}(H_0 \mid x, G_C)$ |
|---|---|---|---|---|---|
| .50 | 1.000 | .894 | .472 | 1.000 | .500 |
| .20 | 3.080 | .351 | .260 | .588 | .370 |
| .10 | 6.314 | .154 | .133 | .309 | .236 |
| .05 | 12.706 | .069 | .064 | .156 | .135 |
| .01 | 63.657 | .0115 | .0114 | .031 | .030 |
| .0032 | 200 | .0034 | .0034 | .010 | .010 |

# Where are we at?

How do p-values and Bayesian model selection stack up in terms of:

- Interpretability?
- Ease of use?
  - Theoretical (understanding behavior)
  - Analytical (coming up with a design in practice)
  - Computational (computing what is needed)
- Counterintuitive behavior?
  - Is counterintuitive behavior possible?
  - Is counterintuitive behavior detectable?
  - Is counterintuitive behavior easy to understand?

# Conclusions

Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. Moving to a world beyond "$p < 0.05$". *The American Statistician*, 73(sup1):1–19, 2019. doi: 10.1080/00031305.2019.1583913. URL https://doi.org/10.1080/00031305.2019.1583913.